Volume 282, number 1, 26–31 FEBS 09612 April 1991

ADONIS 001457939100321E

Structural symmetry of the extracellular domain of the Cytokine/Growth hormone/Prolactin receptor family and Interferon receptors revealed by Hydrophobic Cluster Analysis

E. Thoreau[1], B. Petridou[2], P.A. Kelly[3], J. Djiane[2] and J.P. Mornon[1]

[1]*Groupe Cristallographie et Simulations Interactives des Macromolécules biologiques, Laboratoire de Mineralogie-Cristallographie, Universités Paris VI et Paris VII, CNRS URA 09, T16, 4 place Jussieu, 75252 Paris Cedex 05, France,* [2]*Unité d'Endocrinologie Moléculaire, Batiment des Biotechnologies, INRA, 78350 Jouy en Josas, France* and [3]*Laboratory of Molecular Endocrinology, McGill University, Royal Victoria Hospital, 687 Pine Ave West, Montreal H3A 1A1, Canada*

Received 15 February 1991

Sequence comparison based on Hydrophobic Cluster Analysis procedures shows that the extracellular ≈200 amino acids domains of cytokines receptors belonging to the Cytokine/Growth hormone/Prolactin receptor family and to the Interferon one are organized in two homologous subdomains. Further, comparison of the subdomains of 32 independent sequences and of a lot of already recognized homologous domains with data bases could lead to the hypothesis that these ≈100 amino acids subdomains could possess the overall fold of the constant immunoglobulin domains and so could belong to the immunoglobulin superfamily.

Cytokine; Protein sequence comparison; Protein structure prediction; Immunoglobulin superfamily; Fibronectin; Contactin

1. INTRODUCTION

The rapidly expanding family of the Cytokine/Growth hormone/Prolactin receptors (CRs) now comprises 10 members: IL2Rβ, IL3R, IL4R, IL6R, IL7R, granulocyte-macrophage colony stimulating factor (GMCSFR), granulocyte colony stimulating factor (GCSFR), prolactin (PRLR), growth hormone (GHR), and erythropoietin (EPOR) [1–3]. These receptors possess a single hydrophobic transmembrane area, a highly variable cytoplasmic domain – both in length and sequence – and an extracellular domain with a stretch of about 200 amino acids, here named D200, which is significantly conserved, although at a relatively low level (15–35%) [1]. IL3R has a tandem of two such domains. For GCSFR, three fibronectin type III domains have been recognized between the D200 domain and the transmembrane segment. One additive segment occurs at the N-terminal part of IL6R and of GCSFR and has been recognized to be immunoglobulin like [2,4]

On the other hand, Interferon Receptors (INFRs) show similar organization, either with a unique D200-like domain for the γ (INFgR) or two for the α/β (INFaR). For these receptors it was recently shown [5] that the two D200 domains of INFaR are similar to each other and possess also homology with that of INFgR. Moreover, a structural relationship of the INFRs has been suggested with the Cytokine/Growth hormone/Prolactin receptor family (CR family) [6].

The recognition of conserved features for these biologically related proteins could help the understanding of their mechanisms of action and might open perspectives for therapeutic applications.

In this paper we describe a two-dimensional (2D) sequence analysis of the extracellular hormone binding domain of these receptors. We show that the D200 amino acid conserved domains possess internal symmetry and so probably result from the duplication of a subdomain ≈100 amino acids long, here named SD100A and SD100B. Further, we show that GMCSFR comprises an additional SD100 subdomain at its N-terminal part. These subdomains appear similar to and have the same size as the highly repeated type III domains of fibronectin, which has already been shown to be sequence related to the here-named SD100B [7]. We also confirm and expand the homology with repeated domains of contactin [8].

2. MATERIALS AND METHODS

The protein sequences of all representative members of the CR family and INFRs as well as those of numerous other sequences have been analyzed and compared through Hydrophobic Cluster Analysis (HCA) [9]. This sensitive method is able to detect the similarity of the secondary and tertiary folding of globular protein domains even if their sequence identity is low (<10%) and so, below the limits of effi-

Correspondence address: E. Thoreau, Groupe Cristallographie et Simulations Interactives des Macromolécules biologiques, Laboratoire de Mineralogie-Cristallographie, Universités Paris VI et Paris VII, CNRS URA 09, T16, 4 place Jussieu, 75252 Paris Cedex 05, France

Published by Elsevier Science Publishers B.V.

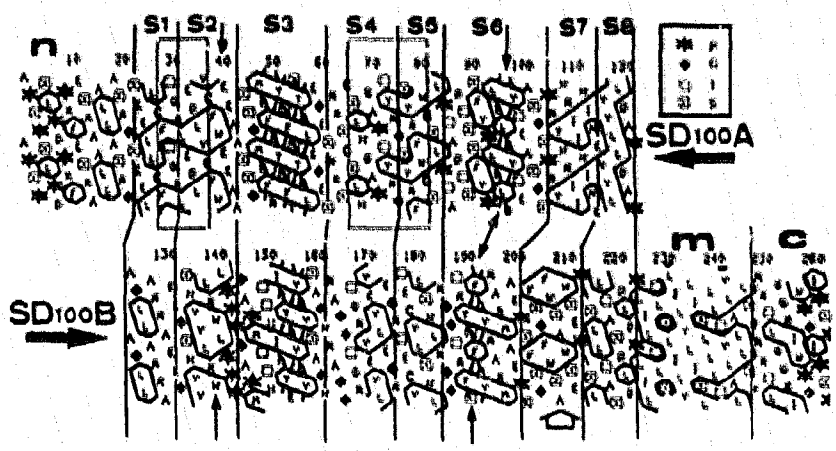


Fig. 1. HCA plots of the extracellular domain of the murine erythropoietin receptor, splitted to show the internal symmetry between SD100A and SD100B subdomains. The N-terminal region, the transmembrane stretch and the beginning of the cytoplasmic domain are respectively indicated by n, m and c. Heavy vertical lines delineate the consensus S1 to S8 structural segments observed in all receptors. Arrows indicate in S2 and in S6 particular conserved features (see text). The large open arrow indicates the WSXWS marker. The two putative S–S bridges are indicated within SD100A by thin lines. HCA score and sequence identity between SD100A and SD100B are 72% and 13%, respectively.

ciency of conventional 1D sequence comparison. A review of the HCA strategies and applications can be found in [10].

Briefly, it appears important to emphasize that the 2D HCA approach primarily relies on the comparison of the hydrophobic amino acid patterns of protein sequences – which are reliable signatures of the 2D and 3D folding of protein domains – rather than only on the maximization principle of identity (or homology) between amino acids which supports classical 1D methods. The sequences were plotted on classical α-helix spreads into a 2D planar pattern ([9,10] and refs therein). The one letter amino acid code is used with the exception of P, G, S and T which are represented by special symbols to help their immediate visual recognitions. P which is a major interrupter of secondary structure (due to its strong internal constraints) is represented by a star, G which contrarily brings a high degree of freedom to the polypeptide chain, is symbolised by a diamond. T and S, which have special hydrophobic mimetic behavior and which are major components of loops, are represented by open and dotted squares, respectively (cf. Fig. 1).

Meanwhile, when necessary and particularly to statistically assess several results of HCA, 1D sequence programs and procedures have been used as described in [10]. The 2D HCA comparisons are, after analysis, currently reported on classical linear alignment through a polyvalent sequence editor (EDITSEQ) developed for this purpose [10]. 3D studies were performed with the MANOSK software [11] running on an Evans & Sutherland PS390 display.

3. RESULTS

Visual inspection of HCA plots of the D200 conserved extracellular domains of CRs and INFRs families led to the frequent observation of clear cut loop (or hinge) regions in their middle (EPDPP for rbPRLR, QPDPP for huGHR, QPPPPKD for muIL3R first D200 domain, KPLAPD for muIL4R, QPDPPAN for huIL6R, GPPE for huINFaR first D200 domain) (see the linear alignment reported in Fig. 3A) similar to that noted between D200 domains, e.g. PPPEN for huINFaR [5]. So a major interrupter of folding could be suspected to be present. Further, HCA plots of many receptors suggest (qualitatively and quantitatively) a structural homology between these two equivalent sized halfs of the D200 domains. Fig. 1 examplifies this with the muEPOR where the middle breaker is DAPAG around position 120. For the two halves of the D200 domain (SD100A and SD100B) of that receptor, HCA-score and the resultant alignment sequence identity are 72% and 13%, respectively. These values are within the range of significant homology otherwise observed for

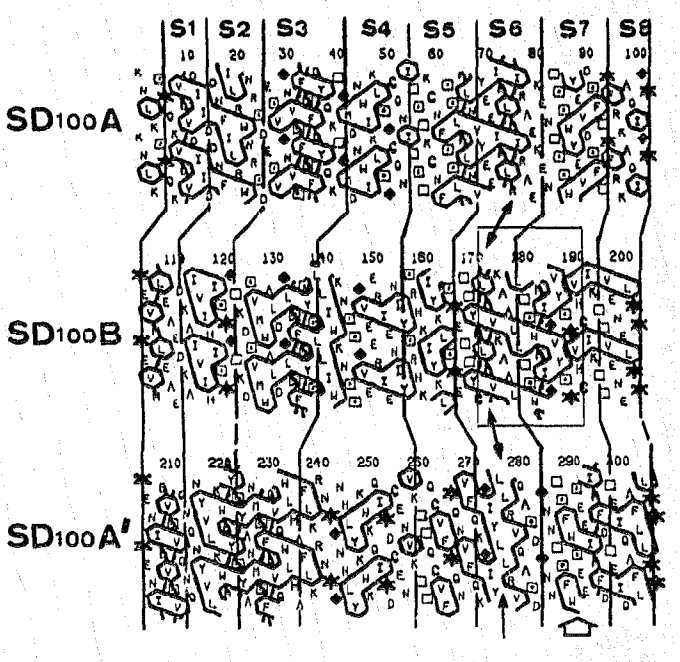


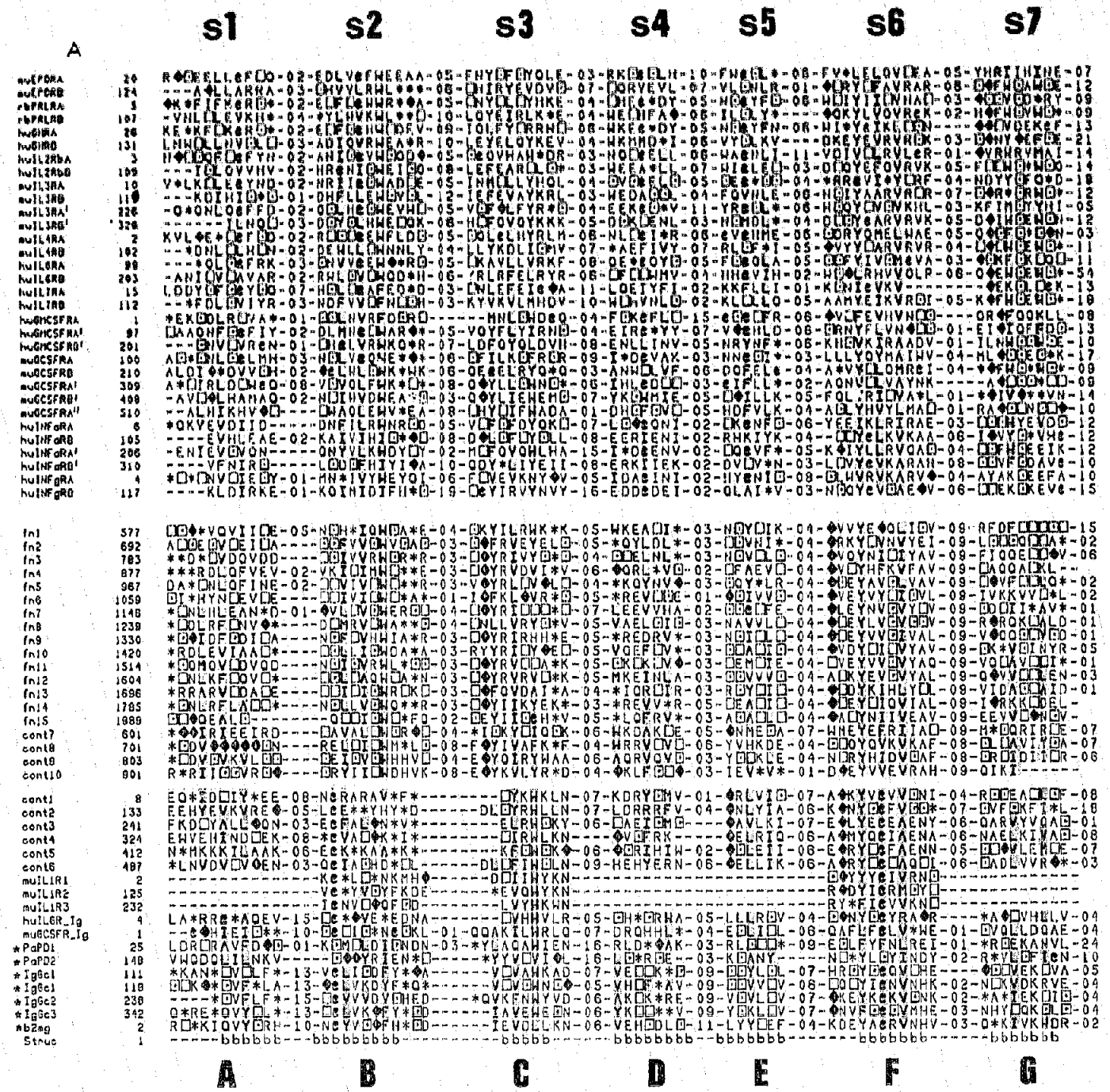
Fig. 2. HCA plots of the first SD100 subdomains of huINFaR, same convention as in Fig. 1. Note that as SD100B of Fig. 1, the third SD100 possesses a FWS motif (large arrow) within the S7 structural segment. HCA scores and sequence identities are (72%, 20%), (67%, 14%), (64%, 16%) respectively for AA', AB and BA' comparisons.

structurally related globular domains of sufficient size [9,10].

Fig. 2 reports the HCA comparison of the SD100A and SD100B of the huINFaR first D200 domain and the SD100A of the second D200 domain. A total of more than 30 half D200 domains, i.e. SD100 subdomains have been analyzed so far.

As already reported [9,10], the HCA handling of a large number of sequences for a family showing medium or low sequence identities considerably reinforce the hypothesis first recognized for particular pairs and leads to certainty. In the present case, both the segmentation of internal parts of domains in loops and hydrophobic clusters (i.e. in more than 8 cases out of 10 of regular secondary structures [10]), and the highly conserved amino acid patterns within nearly all SD100 subdomains (Fig. 3B) provide good evidence for an internal duplication of D200 domains.

The internal symmetry of D200 domains has been further verified by statistical assessment with the



(Fig. 3, for legend see next page.

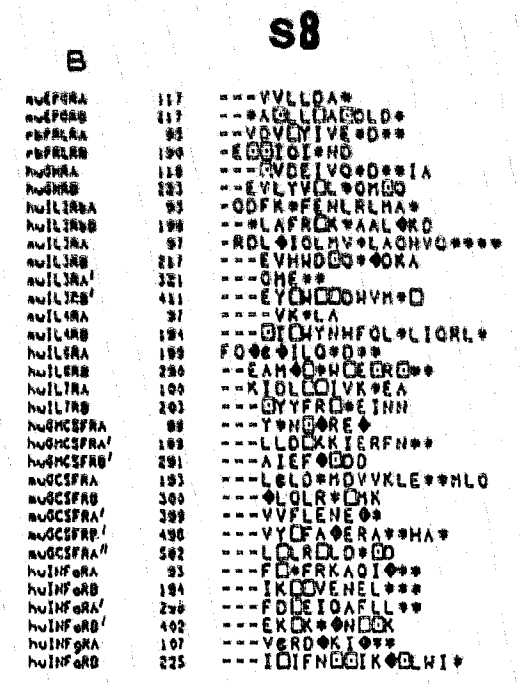


Fig. 3. Overall alignment of SD100 subdomains deduced from 2D HCA plots. Same symbols and labels as in Figs 1, 2. For each sequence, the numbering of the first shown amino acid is indicated; shaded amino acids represent anchors of the alignment. For clarity, variable regions (loops, N or C regions) are not shown; numbers between conserved regions indicate the length of these parts of the sequence; numbers at the end of SD100 indicate the numbers of amino acids to the next SD100 or to the transmembrane segment. From N-terminal to transmembrane segment, SD100 are labeled A, B and A', B' for duplicated D200, the 3 GMCSFR SD100 are labeled A, A', B' and the 5 GCSFR SD100 are labeled A, B, A', B', A". (A) from top to bottom, S1 to S7 segments of CRs, INFRs, fibronectin, Contactin and Ig sequences are shown for comparison. Known 3D structures are prefixed with a star. Last two lines: consensus regions of experimentally observed β-strands in the Ig domains and their current labeling (A to G). (B) S8 and hinge regions of CRs and INFRs SD100.

ALIGN program [12]. A test set of all SD100 comparisons (12 SD100, 66 pairs in all) for muEPOR, muIL3R, huIL7R, huINFaR shows that among previously known corresponding sequences (SD100A/SD100A and SD100B/SD100B) 16 pairs have scores greater than 3.0 standard deviations (SD), considered as significant, the maximum score being 9.51 SD. Among the SD100A/SD100B pairs, the maximum score remains high (6.69 SD for SD100B of the first D200 of muIL3R vs SD100A of the second D200 of huINFaR) and 7 pairs are above 3.0 SD.

All known receptors have 2 or 4 SD100 subdomains, with the exceptions of GCSFR which has 5 SD100 between an immunolike domain [2], residues 98–599, and the transmembrane stretch (one true D200 + three previously recognized fibronectin type III repeats [2]) and GMCSFR which possesses an additional SD100 between the N-terminal and its D200 domain (see Fig. 3A).

3.1. *Comparison of SD100 with fibronectin type III repeats*

Recently, it has been shown that a part of the D200 domains of the CRs family possess a significant sequence homology with numerous repeats encountered within fibronectin, contactin and other proteins [7]. This ≈ 100 amino acid long area corresponds to the second half of the D200 domain, or to the SD100B subdomain.

HCA comparison of these repeats with the SD100 shows that this recognized homology could be extended at the structural 2D and 3D levels to all SD100 of CR and INFR families. Fibronectin possesses three different kinds of repeats [13], all the 15 type III domains are clearly related to CRs and INFRs SD100 (HCA data not shown; see Fig. 3A). For Contactin, HCA defines 4 true SD100 domains (residues 582–962) and 6 preceding SD100-like domains (Fig. 3A: cont7–cont10 and cont1–cont6, respectively).

3.2. *Comparison of SD100 with data bases*

The SD100A of nearly all D200 of CRs shows a pattern of four conserved cysteines which at least for GHR have been shown to exist as 2 successive S–S bridges [14]. SD100A subdomains of INFRs have two conserved cysteines positioned close to the second cysteine pair in the SD100A of CRs. The SD100B of CRs does not possess any conserved cysteines but does have a WSXWS sequence which constitute a clear marker [7]. The SD100B subdomains of INFRs and the SD100a of IL7R, possess another tandem of conserved cysteines at their C-terminal side, the location of which is close to the third S–S bridge in the SD100A of GHR [14]. These consistent particularities may help verify for putative structural homologies which could be detected by HCA within data bases following the currently used procedures [10]. As HCA still necessitates human decisions, it is not yet possible to directly perform a screening of sequence data bases. Thus a large number of sequences potentially related to SD100 were selected with the 1D FASTP program [15] and then 2D analysed through HCA. Many were rejected but a few were conserved for further analysis (see [10] for detailed methodology). Since the number of clearly related SD100 subdomains is high while their sequences are essentially different, a large amount of independent data has been analyzed. Repeatedly, immunoglobulin domains were detected by the 1D process and confirmed by HCA. So it appears that the consensus of the first seven hydrophobic clusters which result from the comparison of all SD100 (Figs 1, 2, 3) could match with the seven conserved β-strands of the immunoglobulin constant domains [16]. The last variable segment S8 (Fig. 3B) could be a link between SD100s and SD100 and the transmembrane segment. Moreover, the shape of hydrophobic clusters of CR as that of INFRs [5] are highly consistent with β-strands [10].

However, if the internal structural symmetry of D200 domains of CR and INFR families is clear, the hypothesis of SD100 belonging to the immunoglobulin superfamily needs further verification. Several structural features support this assumption and are discussed below.

4. DISCUSSION

Tentative alignment of SD100 (CR/INFR/fibronectin type III repeats) with Ig constant domains reported in Fig. 3A led to the following remarks: (i) the general distribution between hydrophobic clusters (essentially regular secondary structures) and loops is as similar between the families as is the distribution of hydrophobic and hydrophilic residues within each segment; (ii) within each family the S2 (B) and S6 (F) segments are more conserved in Ig domains than B and F segments are in the central part of the two β sheets; (iii) within S6 (F), the YX(X')XV motif where X' is an hydrophobic residue, is very frequent; it constitutes a clearly recognized marker of the constant domain of the immunoglobulin superfamily [16]; (iv) the positions of conserved Cys residues of CR and INFR families in the Fc 3D template [17], deduced from the present study are all consistent with the formation of disulphide bridges (paper in preparation) as that first observed for the Ig fold of the chaperone PapD protein [18].

The immunoglobulin fold appears to be a typical example of 3D folds compatible with a large variety of sequences [19,20] even when canonical Cys, Trp residues are absent [18]. This widespread occurrence could be related to the robustness of the fold and/or to evolutionary processes [16]. That the SD100 and thus the D200 belong to this huge superfamily is consequently not as surprising as it might appear. If this hypothesis is further verified for all the SD100s, the previously recognized Ig-like domains of CR (and the related adhesions domains), as well as those of the IL1 receptor [21] (Fig. 3A) constitute a large pool of generally related 3D structures, but specifically designed to allow sophisticated functional abilities.

According to this, the boomerang shape of Fc [17], PapD [18] and part of the human class I histocompatibility antigen, HLA-A2 [22] domains may provide the first overall look of the extracellular domains of these receptor families.

5. LATE NOTE

At the completion of this manuscript an article and a letter from J.F. Bazan [23,24] report the same tentative overall conclusions. As J.F. Bazan used different methodologies, the two approaches reinforce each other. In the above present discussion we confirm by 3D checking the consistency of CRs and INFRs S–S bridges potentialities within the constant Ig fold, which in our opinion is further supported by the YX(V/C)XV consensus sequence of (S6/F) (Fig. 3A). J.F. Bazan [24] considers that IL7R is a unique exception to the domain's receptors architecture. 2D HCA analysis does not support this. IL7R SD100A and SD100B look like each other (Fig. 3A) and IL7R SD100A possesses structural similarities with many other SD100. However, unlike most CRs, SD100A, IL7R SD100A lacks the second putative S–S bridge, but possesses another cysteine pair positioned as INFRs SD100B and GHR SD100A third S–S bridge (Fig. 3A). Our proposition differs from that of Bazan [24] in several locations, mainly:

in S1: huGHRB (5 aa), muIL3RB' (3 aa), muGCSFRB (6 aa), huINFaRA (2 aa)
in S2: huINFaRB' (2 aa)
in S3: huGHRA (2 aa), muIL4RB (5 aa), muGCSFRA (3 aa), huINFaRA (4 aa), huINFaRB (4 aa), huINFaRB' (4 aa), huINFgRA (4 aa), huINFgRB (4 aa)
in S4: huIL2RbB (3 aa), muIL3RB (3 aa), muIL4RB (5 aa), huINFgRB (8 aa)
in S5: muEPORB (4 aa), huGHRB (5 aa), muIL3RA (11 aa), huGMCSFRB' (3 aa), muGCSFRB (2 aa), huINFaRB' (3 aa), huINFgRB (3 aa)
in S6: huIL2RbA (5 aa), muIL3RA (14 aa), muGMCSFRB' (4 aa), muGCSFRA (2 aa), huINFgRA (2 aa)

A major difference occurs for the location of the G strand for SD100; we suggest here the S7 segment, Bazan [24] selects the region corresponding to the S8 one. For this variable region the debate is still open and could possibly be clarified only with new data (sequences and/or 3D ones, including modelling).

These differences may be crucial for further exploitation of the alignments since for extended polypeptide structures, a 3 amino acid error led to a more than 10 Å discrepancy. Alignment based on 2D HCA plots are generally significantly more accurate than 1D ones [10]. However, it should be recalled that in very variable areas, only 3D modelling could make the difference between alternative propositions.

5.1. *Remarks added on revision*

(i) The 3D structure of a human CD4 fragment [25,26] shows an immunoglobulin-like subdomain possessing a non-conventional disulphide bridge similar to those expected to occur in the CR family.

(ii) The sequence of IL5 receptor [27] and IL6 signal transducer gp130 [28] are now known. They are fully compatible with the family and 9 new SD100 subdomains could be added.

Acknowledgements: This study was supported by CNRS, INSERM, INRA, Universities of Paris VI, Paris VII and McGill, Rhône

Poulenc, Institut Scientifique Roussel, Fondation pour la Recherche Médicale and ARC. The authors acknowledge the contribution of L. Lemesle-Varloot.

REFERENCES

[1] Cosman, D., Lyman, S.D., Idzerda, R.L., Beckmann, M.P., Park, L.S., Goodwin, R.G. and March, C.J. (1990) Trends Biochem. Sci. 15, 265–270.
[2] Fukunaga, R., Ishizaka-Ikeda, E., Seto, Y. and Nagata, S. (1990) Cell 61, 341–350.
[3] Bazan, J.F. (1989) Biochem. Biophys. Res. Commun. 164, 788–795.
[4] Yamasaki, K., Taga, T., Hirata, Y., Yawata, H., Kawanishi, Y., Seed, B., Taniguchi, T., Hirano, T. and Kishimoto, T. (1988) Science 241, 825–828.
[5] Gaboriaud, C., Uzé, G., Lutfalla, G. and Mogensen, K. (1990) FEBS Lett. 269, 1–3.
[6] Bazan, J.F. (1990) Cell 61, 753–754.
[7] Patthy, L. (1990) Cell 61, 13–14.
[8] Ranscht, B. and Dours, M.T. (1988) J. Cell Biol. 107, 1561–1573.
[9] Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J.P. (1987) FEBS Lett. 224, 149–155.
[10] Lemesle-Varloot, L., Henrissat, B., Gaboriaud, C., Bissery, V., Morgat, A. and Mornon, J.P. (1990) Biochimie 72, 555–574.
[11] Cherfils, J., Vaney, M.C., Morize, I., Surcouf, E., Colloc'h, N. and Mornon, J.P. (1988) J. Mol. Graph. 6, 155–160.
[12] Dayhoff, M.O., Barker, W.C. and Hunt, L.T. (1983) Methods Enzymol. 91, 524–545.
[13] Skorstengaard, K., Jensen, M.S., Sdhal, P., Petersen, T.E. and Magnusson, S. (1986) Eur. J. Biochem. 161, 441–453.
[14] Fuh, G., Mulkerrin, M.G., Bass, S., McFarland, N., Brochier, M., Bourel, J.H., Light, D.R. and Wells, J.A. (1990) J. Biol. Chem. 265, 3111–3115.
[15] Lipman, D.J. and Pearson, W.R. (1985) Science 227, 1435–1441.
[16] Williams, A.F. and Barclay, A.N. (1988) Annu. Rev. Immunol. 6, 381–405.
[17] Deisenhofer, J. (1981) Biochemistry 20, 2361–2370.
[18] Holmgren, A. and Bränden, C.I. (1989) Nature 342, 248–251.
[19] Amzel, L.M. and Poljak, R.J. (1979) Annu. Rev. Biochem. 48, 961–997.
[20] Pastuszyn, A., Noland, B.J., Bazan, J.F., Fletterick, R.J. and Scallen, T.J. (1987) J. Biol. Chem. 262, 13219–13227.
[21] Sims, J.E., March, C.J., Cosman, D., Widmer, M.B., Robson MacDonald, H., McMahan, J., Grubin, C.E., Wignall, J.M., Jackson, J.L., Call, S.M., Friend, D., Alpert, A.R., Gillis, S., Urdal, D.L. and Dower, S.K. (1988) Science 241, 585–589.
[22] Bjorkman, P.J., Saper, M.A., Samraoui, B., Bennett, W.S., Strominger, J.L. and Wiley, D.C. (1987) Nature 329, 506–512.
[23] Bazan, J.F. (1990) Immunol. Today 11, 350–354.
[24] Bazan, J.F. (1990) Proc. Natl. Acad. Sci. USA 87, 6934–6938.
[25] Wang, J., Yan, Y., Garrett, T.P.J., Liu, J., Rodgers, D.W., Garlick, R.L., Tarr, G.E., Husain, Y., Reinherz, E.L. and Harrison, S.C. (1990) Nature 348, 411–418.
[26] Ryu, S.E., Kwong, P.D., Truneh, A., Porter, T.G., Arthos, J., Rosenberg, M., Dai, X., Xuong, N.H., Axel, R., Sweet, R.W. and Hendrickson, W.A. (1990) Nature 348, 419–426.
[27] Takaki, S., Tominaga, A., Hitoshi, Y., Mita, S., Sonoda, E., Yamaguchi, N. and Takatsu, K. (1990) EMBO J. 9, 4367–4374.
[28] Hibi, M., Murakami, M., Saito, M., Hirano, T., Taga, T. and Kishimoto, T. (1990) Cell 63, 1149–1157.